Modelling and Design of a 45nm SLC 3D NAND Flash CPLD

Arijit Banerjee and Sergiu Mosanu

University of Virginia, Charlottesville, VA 22904, USA

***Abstract*—With the impending end of Moore's law in 2D silicon, there has been development on "more than Moore" technologies like 3D ICs in z-direction. We can stack silicon wafers to build and connect them with through silicon vias to make 3D ASICs as well as 3D FPGAs. However, this approach has lower manufacturing yield and reliability. Recent research has shown promises to make monolithic 3D flash memories, which has revolutionized storage density in flash memory devices. In this paper, we use the concept of monolithic 3D flash memories to make small scale CPLDs using SLC NAND flash architecture and show that the future of large scale FPGAs using NAND flash architecture is promising. This paper shows circuit modelling of 3D CPLDs using modelled SLC NAND flash bitcell and power, area and performance analysis. We further show the architecture of the 3D CPLDs using NAND flash PLA planes and cross sectional details of the NAND string.**

***Keywords*—3D CPLD, 3D visualization, Vertical 3D logic.**

I. Introduction

Shrinking devices with technology scaling is nearing the molecular dimension of the silicon atom, which heralds the end of Moore's law. It is getting harder to fabricate smaller devices nowadays, and researchers are looking for newer options to overcome the challenges in better design metrics of devices, chips, and systems. As the short channel device properties deteriorates in bulk sub $45nm$ devices, the advent of FINFETs and FDSOI devices provide us with better subthreshold slope, leakage current, and ON currents. However, further device scaling in 2D silicon is getting harder and researchers are looking for novel transistors like carbon based graphene [1], carbon nano-tubes (CNT) [2], group III-VI compound semiconductors [3], tunnel FETs [4], etc. On the memory side, there has been new developments for spin torque transfer (STT) RAMs [5], ferromagnetic RAMs (FRAM) [6], flash multi-level charge trap cells, etc. In addition to the search of the next generation devices, researchers are looking for increasing the logic and memory density by stacking multiple dies on top of each other and going to 3D ICs. This method of stacking multiple dies to go 3D for increasing the logic or memory density requires big fat inter die signal and power connectors called through silicon vias (TSV). However, the TSV based 3D ICs have reliability and yield issues [7] due to various reasons like poor thermal energy extraction, TSV signal-integrity issues, TSV manufacturing faults and break down issues, etc. On the other hand, there are technologies to integrate multiple die in a package which can also solve some of the logic density issues that we are currently facing. These technologies include multiple die on a package [8], dies on silicon or other interposers [9], etc., and are called 2.5D technologies. However, none of these technologies are monolithic in nature and due to the additional assembly cost in stacked-via 3D ICs, the overall cost of the 3D ICs is higher.

Flash memories have been there for a while starting from floating gate MOSFETs (FGMOS). Recent development in NAND flash memories in monolithic 3D direction [10][11][12][13][14][15] has revolutionized the world of storage and products are available from Samsung as 3D-VNAND [15] for purchase nowadays. These monolithic 3D flash technology not only increases the density of the transistors, but also reduces routing length by going 3D and thus the wire length cap is reduced. Consequently, the energy consumption for 3D NAND flash memories is lower than the conventional NAND flash.

Apart from the recent development in devices and novel materials, there has been increased focus on field programmable gate arrays (FPGA). FPGAs have built in flexibility to reconfigure logics that enable us to implement and test products ahead of time. Prototyping a circuit or a system in FPGAs are much lower in cost that making them in application specific integrated circuits (ASIC). Recent development in 3D FPGAs [16] using stacked dies and TSVs have made the logic density even higher with lesser cost per transistor. However, we can achieve much higher logic density than stacked die 3D using monolithic 3D technology used in 3D NAND flash technologies. This paper proposes the use of 3D NAND flash technology for configurable logic blocks (CLB) in complex programmable logic devices (CPLD) to further increase the logic density or programmable literal density. In this paper, we propose the usage of 3D NAND flash arrays as NAND planes to make a CLB and using multiple CLBs to make a CPLD. We further propose that this approach can be extended to build a monolithic 3D FPGA using 3D NAND flash technology. The rest of the paper is divided into several sections. Section II describes the 3D technologies for ASICs, FPGA and flash memories in brief. Section III describes a floating gate model for 3D SLC NAND bitcell. Section IV describes the architecture of the CLB. Section V describes possible architectures of the CPLD and extends it further for FPGAs. Section VI describes the visualization and 3D physical design of the 3D NAND planes in the CPLDs and FPGAs. We conclude in section VII.

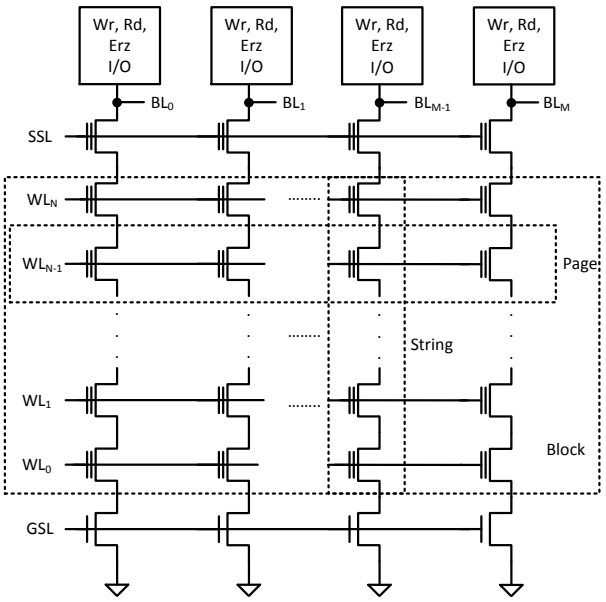

Fig. 1. 2D NAND flash array organization.

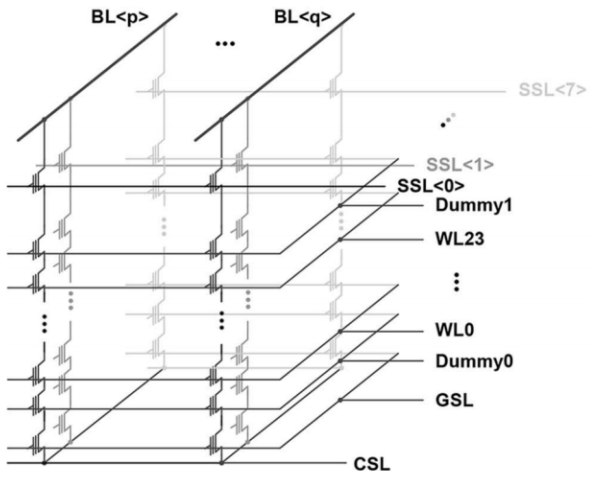

Fig. 2. 3D NAND flash array organization [15].

II. More Than Moore 3D Technologies For ASIC, CPLD, FPGA And Flash Memory

As Moore's law of scaling is nearing an end in 2D scaling in silicon, going 3D is one way to impede the end of Moore's law for a few decades. As discussed earlier, 3D technologies use different methods to create 3D, such as stacked die 3D with TSVs, monolithic 3D technology, etc. The 3D technologies provide us with higher density of transistor packing in the same footprint, and thus increase the logic density drastically. We can use these 3D technologies for application specific integrated circuits (ASIC) [17], field programmable gate array (FPGA) [16] and memories like SRAM [10], ROM or non-volatile flash memories [13]. 3D ASICs are usually build in the stacked die approach using TSVs due to its non-regular structures. However, CPLDs, FPGAs or memories have very regular structures, and we can use monolithic 3D technology to make these 3D CPLDs, FPGAs or memories. CPLDs usually have multiple programming array logic (PLA) AND-OR planes to program complex logic functionality as glue logics for various other ICs. We can also use a NAND-NAND PLA planes in a CPLD for programming. Thus we can replace the programmable NAND planes with 3D NAND flash planes in the 3D CPLDs. In this paper, we use the concept of monolithic 3D NAND flash memories to build regular programmable logic structures in monolithic 3D CPLDs and further extending for the use in monolithic 3D FPGAs.

Fig 1 shows a typical 2D NAND flash memory array organization using SLC NAND bitcell. The NAND array is called a block that comprises pages and strings as shown in Fig 1. The pages are bitcells in the same wordline (WL) and the strings are bitcells in the same bitline (BL). The read, write and erase operation is being supported by applying non-standard supply voltages in WL and BLs using the the wordline driver (not shown in Fig 1) and the Wr-Rd-Erz I/Os (Fig 1). We use the GSL lines for selecting a block for read, write and erase operations. Fig 2 shows a typical 3D NAND array organization. From the circuit-topology standpoint the 3D NAND array is very close to the 2D NAND array organization. The concept of page, string and block remains the same. However, in 3D NAND array, the strings are physically located as vertical pillars along the Z-axis, and there can be multiple array planes clubbed together in an array block. We manage the strings using the string select lines (SSL) that uses string select decoder (not shown in Fig 2). There are multiple planes of arrays in a 3D NAND block and each plane shares a single bitline. Note that the SSL lines select the same strings in different array planes and we share the wordlines (Fig 2) among a 2D page in the 3D NAND flash array. In addition, we use the GSLs in 3D NAND the same way to perform a read, write or erase operations like in 2D NAND flashes.

III. Floating gate MOS model for SLC bitcell

In order to model the behaviour of an SLC NAND bitcell that behaves like a floating gate MOSFET (FGMOS), we use a Veriloga model to mimic the read, write and erase behaviour of a FGMOS that operates with the Fowler Nordheim tunnelling (FNT) principle. Fig 3 (a) shows the full model for the SLC FGMOS bitcell. Apart for the Veriloga model, we use a voltage controlled voltage source (VCVS) to generate a voltage $V(N)$ that can apply a gate voltage to make the ordinary NMOS (Fig 3 (a)) a depletion type if we do a write '1' operation. On the other hand, the $V(N)$ would be $0V$ for a erase or stored '0' condition for the SLC bitcell that will behave as a normal enhancement type NMOS. Note that the Veriloga model senses the gate-source voltage (V_{GS}) to implement the write '1' operation. Similarly, the Veriloga model senses the gate-body voltage (V_{GB}) of the NMOS transistor (Fig 3 (a)) to implement the erase operation in the bitcell. We program the Veriloga model in such way that if the V_{GS} overshoots $1.7V$ it performs a write '1' operation by creating a $1.7V$ of $V(N)$. On the other hand, for erase operation, we program the model

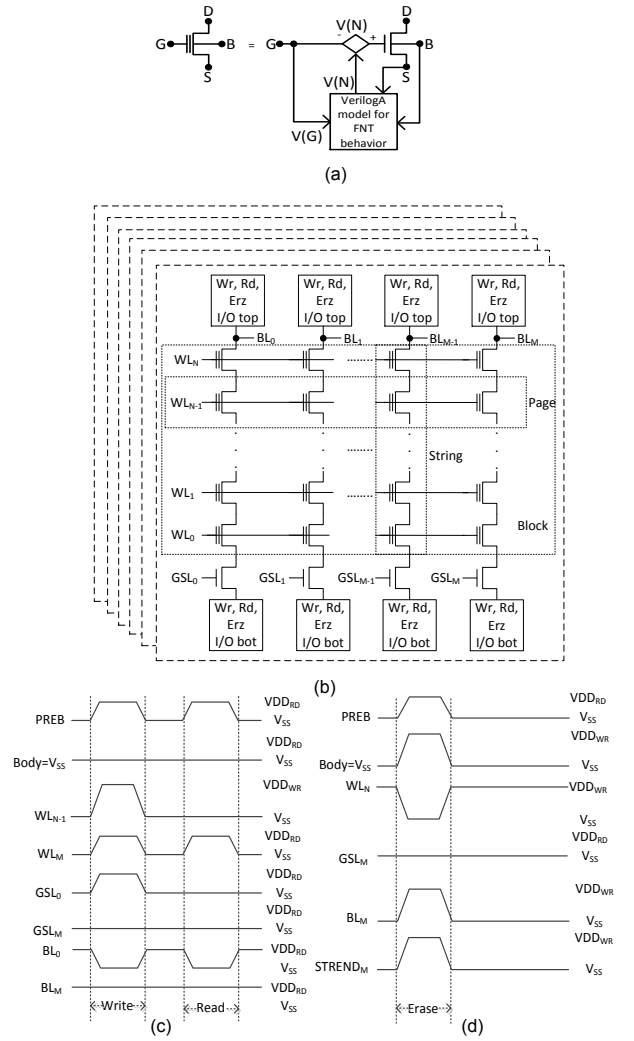

Fig. 3. (a) FGMOS model as a SLC 3D NAND bitcell, (b) 3D NAND array organization using NAND SLC FGMOS bitcell for 3D CPLD and FPGAs, (c) write and read waveforms, (d) erase waveform.

to erase if the V_{GB} undershoots $-1.5V$ and it creates $0V$ of $V(N)$.

CPLDs usually have programmable PLA AND-OR planes. For 3D CPLDs we apropose programmable NAND-NAND plane using 3D NAND arrays. In order to use the NAND flash as programmable planes in the 3D CPLD or FPGAs, the 3D array organization of the 3D NANDs needs modification. Fig 3 (b) shows the modified 3D NAND array organization for 3D CPLDs and FPGAs. Note that we connect the bitlines to separate strings in a block and wordlines are separate for each plane and for each page. This architecture of WL and BLs for 3D CPLD array organization is necessary to support programming of literals and connecting them to the next programming planes in the CPLD. Fig 3 (c) and Fig 3 (d) show the write and read waveforms for the 3D NAND planes in the CPLDs. Here, we assume that we use only two non-standard supply voltages for read, write and erase operations as $VDD_{RD} = 1.2V$ and $VDD_{WR} = 2.2V$. Note that we use NCSU $45nm$ PDK for designing the CLB circuit as well as architecture in Cadence Virtuoso, which internally uses predictive technology model (PTM) from ASU.

IV. Architecture Of The CLB Using 3D NAND For 3D CPLD And FPGAs

In order to support programmability in 3D CPLDs, we need to use multiple configurable logic blocks (CLB), that use two 3D NAND flash programmable planes in each CLB. Fig 4 shows our proposed architecture for the CLBs using two 3D NAND planes. The CLB has two modes for users as program mode and run mode. In the program mode the user needs to program and configure the CLB for a mapped logic functionality. The logic mapping or programming can be done using a custom tool flow or free tools like Verilog-to-Routing Project (VPR). As per our array organization, we allow a maximum of 8 literals per array planes (16 wordlines as equivalent) those can interact with each other to generate independent 8 output from the first plane (Fig 4). The second plane can further extend the complexity of the logic by creating more complex logic functions using the derived outputs of the first plane to produce further 8-bit output from the second plane. Note that the read operation of the 3D NANDs is implemented using dynamic logic. Therefore, we need some latches to hold the data even after the read is done. The literal bus driver (Lbus Driver) transforms the 8-bit literal to 16-bit wordlines towards the first 3D NAND plane. Similarly, the 8-bit output of the first NAND plane after being latched uses another Lbus Driver to generate the wordlines for the second 3D NAND plane. We use an address program decoder that selects the appropriate wordlines in the 3D NAND planes for user programming. The output the second stage of the 3D NAND plane after the latches goes to the filflops and multiplexer at the same time to get selected for operations as sequential logic or combinatorial logic. However, due to the use of dynamic logic in 3D NAND flash planes the literals cannot be truly combinatorial and thus the latch clock needs to be updated in a timely fashion. We implement the the read, write and erase operations for the 3D NAND planes using the top and the bot I/O (Fig 4). Fig 5 shows the circuits implementation for the top and bot I/O. We generate the non-standard supply voltages for the wordlines using the programming wordline decoder that has wordline drivers to supply the non-standard voltages to the wordlines as shown in Fig 6. Note that the typical read, write and erase voltages at the time of operation are the same as described earlier; however, we optimize the wordline waveforms for read, write and erase to toggle less and they look different that the waveforms shown in Fig 3.(c) and Fig 3.(d). Note that we do not design the programming state-machine of the CLB due to time constraints; however, we define all the necessary pins to operate the CLB using a synthesized state-machine using a Verilog or VHDL description.

V. Architecture Of The CPLD And FPGAs Using 3D NAND CLBs

In general, an FPGA fabric or architecture consist of multiple CLBs that connects to each other as per user definition

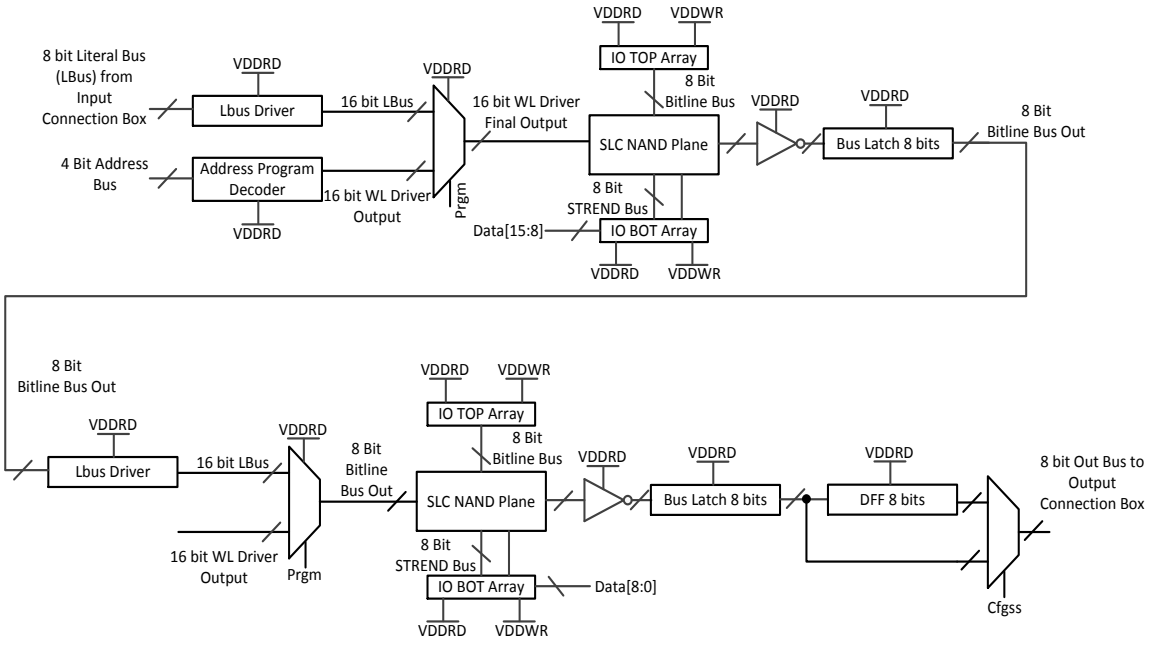

Fig. 4. Architecture for 3D NAND based CLBs for CPLD and FPGAs

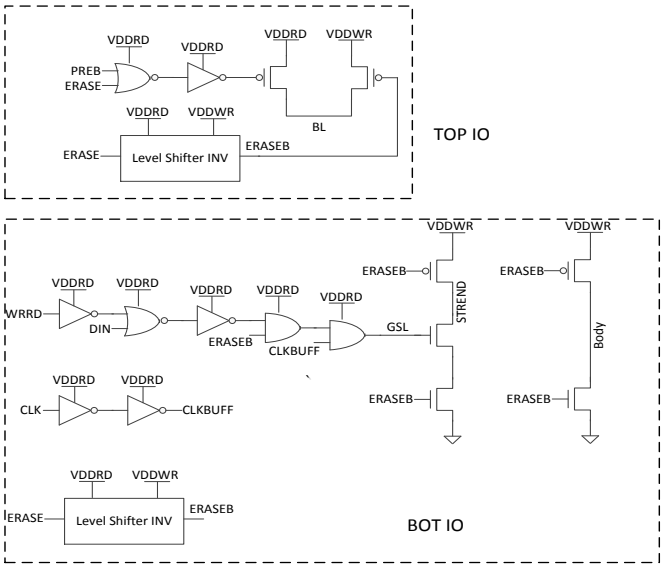

Fig. 5. Top and bot I/O for 3D NAND SLC bitcell array supporting read, write and erase operation.

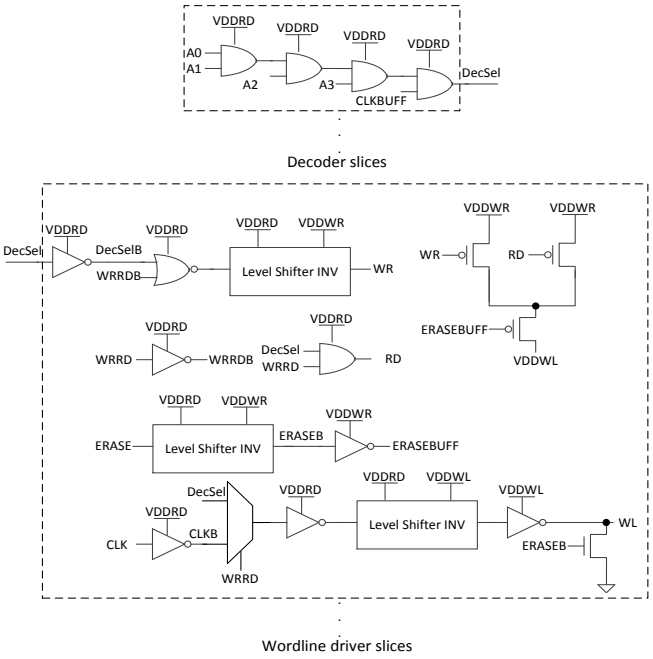

Fig. 6. Wordline driver for SLC 3D NAND flash array to support read, write and erase operation.

using input and output connection boxes [18]. The signals from and to the CLBs, route through the available routing tracks using a switch-box topology [18]. We propose to us the CLB described in the previous section in a FPGA type fabric to make a CPLD. The architecture of the 3D CPLD using 3D NAND flash can be of two types. In one type, the 3D NAND flash can be physically located inside each CLB blocks. We name it as distributed 3D CPLD architecture as shown in Fig 7. On the other type of architecture, the 3D NAND planes can be centralized and located in the centre of the CPLD die (Fig 8). Both architectures have some benefits and drawbacks. The distributed architecture has lower routing cost in terms of WL and BL routability; however, there can be mask cost overhead and more structural brittleness in this architecture. On the other hand, the centralized architecture has lower mask cost and lesser structural brittleness, but suffers from routing congestion and need dedicated WL and BL routing to and from CLBs to

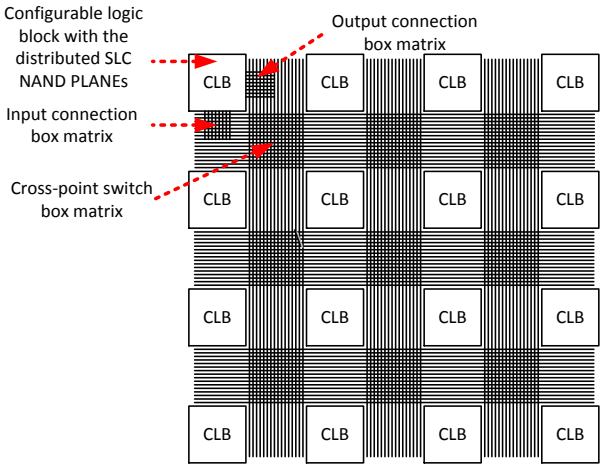

Fig. 7. CPLD architecture for distributed 3D NAND planes.

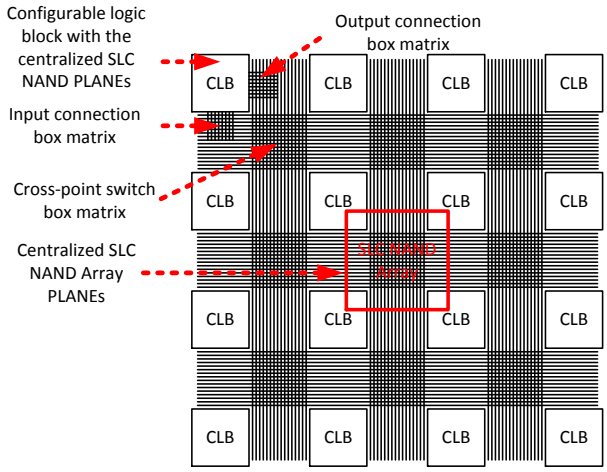

Fig. 8. CPLD architecture for centralized 3D NAND planes.

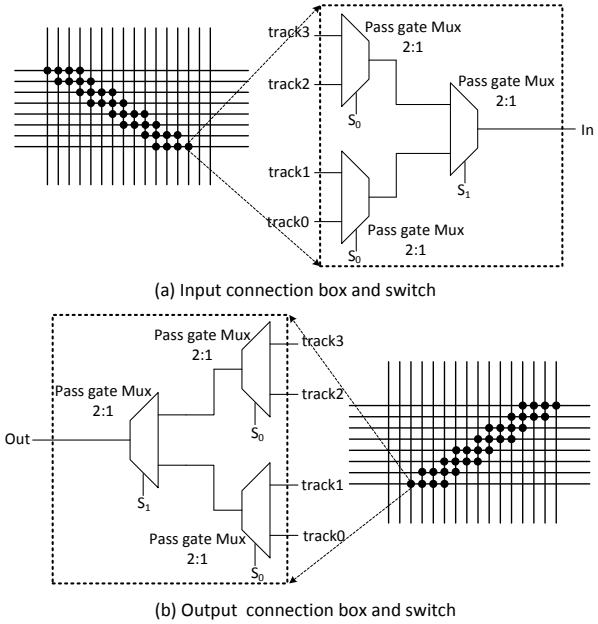

Fig. 9. Topology for input and output connection boxes in 3D CPLD and FPGAs.

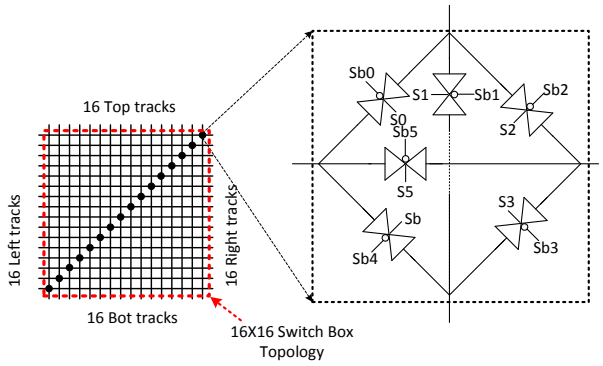

Fig. 10. Topology for switch boxes in 3D CPLD and FPGAs.

the centralized 3D NAND planes. The routing fabric of our 3D CPLD has 16-bit routes as shown in Fig 7 that has routing length of one unit as we place switch boxes in every vertical and horizontal crossings. This CPLD architecture has 16 total CLBs that connect to the main routing channels using 8-bit input and output connection boxes. Fig 9 shows the topology and actual circuit of the input and output connection boxes. Note that the flexibility of the input and output connection box is four. The 1-bit input and output connection box uses PMOS-NMOS transmission gates for connection. Fig 10 shows the topology of the switch box used in our 3D CPLD. This switch box topology is much simpler than the planer (2D) [19] and Wilton topology [19] and using the connection box flexibility of four our 3D CPLD architecture can support flexible routing options. Note that we use PMOS-NMOS transmission gates to connect routes in the switch box. To configure the connection box and the switch box, one can program the associated SRAM bits for the transmission gates accordingly, which is not shown in Fig 9 and Fig 10.

In order to estimate the literal density of the 3D NAND flash CPLD, we assume that the reference CPLD uses a 2D version of the NAND flash and the effective area of the 2D NAND plane in XY surface are same as the 3D NAND planes area in the XZ surface. We found that using 16 layers of 3D NAND stacks in the 3D CPLDs, we have approximately $0.48\mu m^2$/literal for 2D NAND flash CPLD while it is $0.045\mu m^2$/literal for 3D NAND flash CPLD which is $10\times$ more denser than the 2D one. We report the average write power is $44.65mW$ per literal, read power is $51.53mW$ per literal and erase power per block is $66.49mW$ for $495MHz$ of program clock frequency.

We can further extend this CPLD architecture to big monolithic 3D FPGAs by using 10K or more CLBs in the architecture propose in Fig 7 and Fig 8. For big 3D FPGAs, we may need complex topology of the connection boxes and switch boxes. In addition, we need to increase the number of track routes to 200 or more to support complex routability that requires different length of routes.

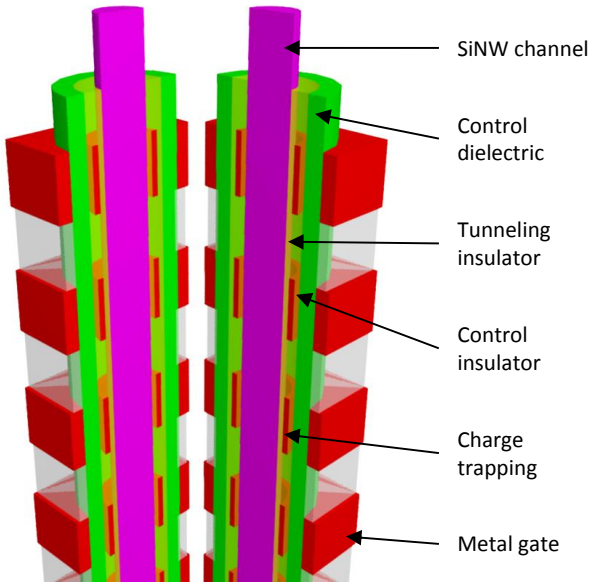

Fig. 11. Cross-section layer structure of a NAND string in cylindrical shape.

VI. Visualisation And Physical Design In 3D

For 3D visualization, we use Autodesk Maya and develop a Maya Embedded Language (MEL) script to generate the 3D NAND chip structure. We model the NAND planes as 16×8 FGMOS in cylindrical form, forming an array of 8 string NAND pipes per NAND plane in the CPLD.

A. NAND string layer structure

The innermost cylinder represents the silicon nanowire (SiNW) channel, with the source connector at the bottom and the drain connector at the top. The channel is surrounded by an insulator layer, which fully encloses inside 16 charge trapping all-around metal floating gates. Tunnelling will occur between the channel and the floating gates through the insulator during the write operation in the transistor. Following is a dielectric layer, which is in contact with 16 all-around gates. The dielectric layer will provide additional insulation and better control between the gate and the floating gate. Since the channel is a SiNW, the drain of a transistor connects with the source of the upper transistor; therefore, the transistors are connected in series in a NAND configuration. The insulator and dielectric of all the transistors together shall therefore resemble a one-body pipe structure, as depicted in Fig. 11.

B. Bitline and wordline wiring

Contrary to Flash memory, the CPLD circuit requires that each bitline output is wired to the planar silicon logic, which is challenging, since the number of wires increases considerably with increase in the size of the NAND plane. A maximum utilization of area can be achieved when the distance between 2 neighbouring NAND pipes in a plane and the distance between 2 neighbouring NAND planes are equal. For that, the bitline

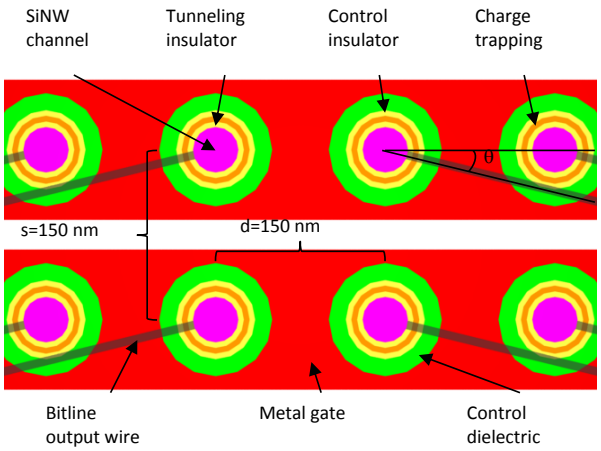

Fig. 12. Top view depicting cell dimensions and bitline wires placement

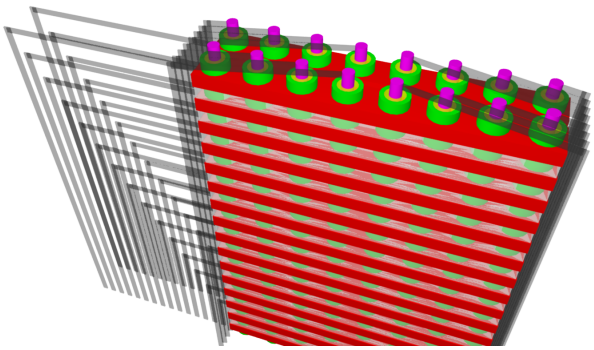

Fig. 13. CLB consisting of 2 8×16-NAND planes, with separate wires for each wordline and bitline.

wires need to be very thin and placed on top of the dielectric pipe, connected only to the top of the SiNW at an angle θ with the plane, as shown in Fig. 12. In our design, 8-bitlines are connected with the planar silicon logic with 8-wires, 4 to each side, which allows thicker wires and larger spacing between the wires. Moreover, each wordline has to be connected to the planar silicon logic, for which a structure has been designed as shown in Fig. 13. There are several options of connecting the wordlines. Previous literature describes an innovative stair structure. An alternative is to wire 8 wordlines on each side. In our design, all 16 wordlines extend on one side of the plane.

C. From a NAND plane to full CPLD

The NAND plane in this project shall consist of 8×16 transistors, therefore 8 such cylindrical structures are aligned in a row. For NAND-NAND logic, 2 such NAND planes are aligned, forming the two programming planes in a CLB block, as depicted in Fig. 13. The complete CPLD with the centralized 3D NAND architecture consists of 4×4 CLB blocks, as is depicted in Fig. 14. Clearly, the challenging part in 3D logic is managing such a high number of wires that requires further wiring design optimizations.

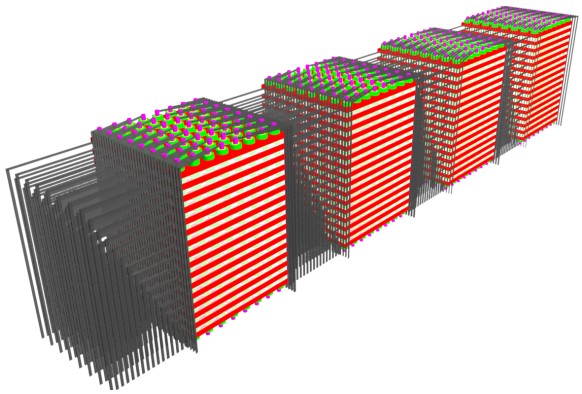

Fig. 14. Full CPLD consisting of 4×4 CLBs.

D. Footprint requirements

A CLB block will contain $16 \times 8 \times 2 = 256$ 3D NAND transistors. If we omit the wiring footprint, a CLB will occupy an area of $(150nm \cdot 2) \times (150nm \cdot 7 + 150nm) = 0.36\mu m^2$, which translates into $0.0225\mu m^2$ per NAND string, or $0.045\mu m^2$ per literal. However, with the current wire configuration, a CLB footprint is $(150nm \cdot 2) \times (2000nm) = 0.6\mu m^2$, or $0.075\mu m^2$ per literal.

VII. Conclusion

This paper shows the first modeling of circuits for 3D CPLDs in a $45nm$ technology. We show that using 3D NAND flash arrays we can make 3D PLA planes in CLBs to make the basic building configurable logic block of 3D CPLD with 8 literals. We further show that this CPLD architecture can be extended to a bigger FPGA architecture. Our CPLD has a $0.045\mu m^2$ per literal which has $10\times$ smaller footprint than 2D flash implementation of the same CPLD. Our CPLD's configurable logic block has average $44.65mW$ of write power, average $66.49mW$ of erase power per block. The average read power consumption of the configurable logic block is $51.53mW$ in our CPLD.

Appendix A
Compute θ

The angle θ can be computed as:

$$\tan^{-1}\frac{r_{channel} + space}{d} < \theta < \tan^{-1}\frac{s - r_{channel} - space}{3 \cdot d}$$

Acknowledgment

We thank Dr. Mircea Stan and ECE 7332 class students for providing their valuable feedback and suggestions to make the project successful.

References

[1] J.S. Moon et al., "Self-aligned graphene-on-SiC and graphene-on-Si MOSFETs on 75 mm wafers,", DRC, 2010, vol., no., pp.209,210, 21-23 June 2010.

[2] J. Cheon et al., "Fabrication of n-Type CNT Field-Effect Transistor Using Energy Band Engineering Layer Between CNT and Electrode,", Electron Device Letters, IEEE, vol.34, no.11, pp.1436,1438, Nov. 2013.

[3] J.A. del Alamo, "InGaAs nanoelectronics: from THz to CMOS,", EDSSC, 2013 IEEE International Conference of, vol., no., pp.1,1, 3-5 June 2013.

[4] A. Mishra et al., "Double gate vertical tunnel FET for hybrid CMOS-TFET based low standby power logic circuits,", Emerging Research Areas and AICERA/ICMiCR, 2013 Annual International Conference on, vol., no., pp.1,4, 4-6 June 2013.

[5] S. Zhenyu et al., "STT-RAM Cache Hierarchy With Multiretention MTJ Designs,", VLSI Systems, IEEE Transactions on, vol.22, no.6, pp.1281,1293, June 2014.

[6] H.H. Kim et al., "Novel integration technologies for highly manufacturable 32 Mb FRAM,", VLSI Technology, 2002. Digest of Technical Papers. 2002 Symposium on, vol., no., pp.210,211, 11-13 June 2002.

[7] J. Xie, Y. Wang, Y. Xie, "Yield-aware time-efficient testing and self-fixing design for TSV-based 3D ICs,", ASP-DAC, 2012 17th, vol., no., pp.738,743, Jan. 30 2012-Feb. 2 2012.

[8] C. Palesko, A. Palesko, E.J. Vardaman, "Cost and yield analysis of multi-die packaging using 2.5D technology compared to fan-out wafer level packaging,", ESTC, 2014, vol., no., pp.1,5, 16-18 Sept. 2014.

[9] Q. Chen et al., "Modeling, Fabrication, and Characterization of Low-Cost and High-Performance Polycrystalline Panel-Based Silicon Interposer With Through Vias and Redistribution Layers,", Components, Packaging and Manufacturing Technology, IEEE Transactions on, vol.4, no.12, pp.2035,2041, Dec. 2014.

[10] O. Thomas et al., "Compact 6T SRAM cell with robust read/write stabilizing design in 45nm Monolithic 3D IC technology,", ICICDT '09. IEEE International Conference on, vol., no., pp.195,198, 18-20 May 2009.

[11] S. Panth, K. Samadi, D. Yang, S.K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology,", ASP-DAC, 2013 18th, vol., no., pp.681,686, 22-25 Jan. 2013.

[12] C.H. Shen et al., "Monolithic 3D chip integrated with 500ns NVM, 3ps logic circuits and SRAM,", IEDM, 2013 IEEE International, vol., no., pp.9.3.1,9.3.4, 9-11 Dec. 2013.

[13] Y. Kim et al., "Three-dimensional NAND Flash architecture design based on single-crystalline STacked ARray," IEEE Transactions on electron devices, Vol. 59, No. 1, January 2012.

[14] B. Prince, "Vertical 3D memory technologies," John Wiley and Sons Ltd 2014.

[15] K.T. Park, et al., "Three-dimensional 128 Gb MLC vertical NAND Flash memory with 24-WL stacked layers and 50 MB/s high-speed programming," IEEE JSSCC, Vol. 50, No. 1, January 2015.

[16] F. Furuta et al., "Scalable 3D-FPGA using wafer-to-wafer TSV interconnect of 15 Tbps/W, 3.3 Tbps/mm2,", VLSIC, 2013 Symposium on, vol., no., pp.C24,C25, 12-14 June 2013.

[17] M.J. Wolf, P. Ramm, A. Klumpp, H. Reichl, "Technologies for 3D wafer level heterogeneous integration,", Design, Test, Integration and Packaging of MEMS/MOEMS, 2008. MEMS/MOEMS 2008. Symposium on, vol., no., pp.123,126, 9-11 April 2008.

[18] V. Betz, J. Rose, A. Marquardt, "Architecture and CAD for Deep-Submicron FPGAs,", Kluwer Academic Publishers, Norwell, MA, 1999.

[19] J.S. Soofiani, N. Masoumi, "Area efficient switch box topologies for 3D FPGAs,", NEWCAS, 2011 IEEE 9th International, vol., no., pp.390,393, 26-29 June 2011.